AFRL-RH-WP-TP-2010-0008

Using a Computer Game for Research on Culture and Team Adaptability: Lessons Learned from a NATO Experiment

Rik Warren
Janet Sutton

Warfighter Interface Division
Cognitive Systems Branch

February 2008

Interim Report

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Cognitive Systems Branch
Wright-Patterson AFB OH 45433-7022

NOTICE

AFRL-RH-WP-TP-2010-0008

//SIGNED//
RIK WARREN
Program Manager
Cognitive Systems Branch

//SIGNED//
DANIEL G. GODDARD
Chief, Warfighter Interface Division
Human Effectiveness Directorate
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
February 2008	Interim	

4. TITLE AND SUBTITLE

Using a Computer Game for Research on Culture and Team Adaptability: Lessons Learned from A NATO Experiment

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)

Rik Warren, Janet Sutton

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
71841009

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Cognitive Systems Branch
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RHCS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-WP-TP-2010-0008

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
NATO-RTO-MP-HFM-142 Symposium "Adaptability in Coalition Teamwork," Copenhagen, Denmark, April 21-23, 2008. 88th ABW/PA cleared on 09 May 2008, WPAFB-08-2836.

14. ABSTRACT

In order to investigate the performance of mixed- versus homogeneous-culture military teams, the NATO RTO Research Task Group, HFM-138/RTG on Adaptability in Multinational Coalitions conducted an experiment using a complex, but very absorbing and immersive, computer-based role-play game using a modern urban search-for-contraband scenario. Game-play required planning, resource allocation, situation awareness, communication, and coordination for successful performance. This paper briefly describes the experiment and its results prior to discussing the lessons learned in conducting the experiment. It focuses on practical methodological and logistical implications for future research on culture and teamwork using computer games in general. It also considers deeper issues in hypothesis generation, scenario and task definition, experimental design, data analysis, and results presentation and communication.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR	12	Rik Warren

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

THIS PAGE LEFT INTENTIONALLY BLANK

Using a Computer Game for Research on Culture and Team Adaptability: Lessons Learned from a NATO Experiment

Rik Warren & Janet Sutton
US Air Force Research Laboratory
Human Effectiveness Directorate
Wright-Patterson AFB OH 45433-7022 USA

Richard.Warren@wpafb.af.mil & Janet.Sutton@wpafb.af.mil

ABSTRACT

In order to investigate the performance of mixed- versus homogeneous-culture military teams, the NATO RTO Research Task Group, HFM-138/RTG on Adaptability in Multinational Coalitions conducted an experiment using a complex, but very absorbing and immersive, computer-based role-play game using a modern urban search-for-contraband scenario. Game-play required planning, resource allocation, situation awareness, communication, and coordination for successful performance. This paper briefly describes the experiment and its results prior to discussing the lessons learned in conducting the experiment. It focuses on practical methodological and logistical implications for future research on culture and teamwork using computer games in general. It also considers deeper issues in hypothesis generation, scenario and task definition, experimental design, data analysis, and results presentation and communication.

1.0 THE NATO RTO HFM-138/RTG COMPUTER GAME EXPERIMENT

Good communications is crucial for (possibly geographically distributed) team members conducting a complex military task such as searching for hidden weapons in an urban environment. It is reasonable to presume that effective communication should be easiest for people who share a common culture. Hence, the NATO Research and Technology Organization (RTO) Human Factors and Medicine Panel Research Task Group on Adaptability in Coalition Teamwork (HFM-138/RTG) conducted an experiment entitled "Leader and Team Adaptability in Multinational Coalitions (LTAMC)" to investigate the performance of mixed- versus homogeneous-culture military teams. Before we can discuss the lessons learned from this experiment, we need to briefly review its hypothesis, methods and principal results. For a more detailed treatment, see NATO RTO HFM-138/RTG (2008).

1.1 Hypothesis & Scenario

The principal hypothesis was that teams whose members are all from the same nation perform better than teams whose members are from different nations. The experiment utilized a complex, but very absorbing and immersive, computer-based role-play game using a modern urban search-for-contraband scenario specifically tailored for this NATO experiment [Leung, Diller, & Ferguson, 2005; Warren, Diller, Leung, Ferguson, & Sutton, 2005] which required planning, resource allocation, situation awareness, communication, and coordination for successful performance. Good performance also required maintaining the good-will of the local "populace" (i.e., computer-generated characters) who could provide useful or misleading information to the search team.

1.2 Participants

The experiment involved 56 four-person teams (224 military officers in all). In 48 of the teams, all four team members were from the same nation; in 8 of the teams, the four members were from different nations. For experimental design purposes, we have 7 national groups of 7 to 9 teams each: Bulgaria (8 teams), The Netherlands (8 teams), Norway-senior (8 teams of senior officers), Norway-junior (8 teams of junior officers), United States (7 teams), Sweden (9 teams), and Mixed nationality (8 teams).

Within each team, the members had to be no more than one rank apart. Although there was no age requirement, the rank constraint meant that team members were of comparable ages. The computer game-play was all in English and all communication was by keyboard. Hence, all participants had to have met a NATO-required level of English proficiency and a reasonable, but unspecified, level of computer experience. Post-play metrics of English proficiency and game experience were developed from responses to pre-game questionnaires. The resulting national, age, English proficiency, and game experience profiles of the 56 teams are shown in Figure 1.

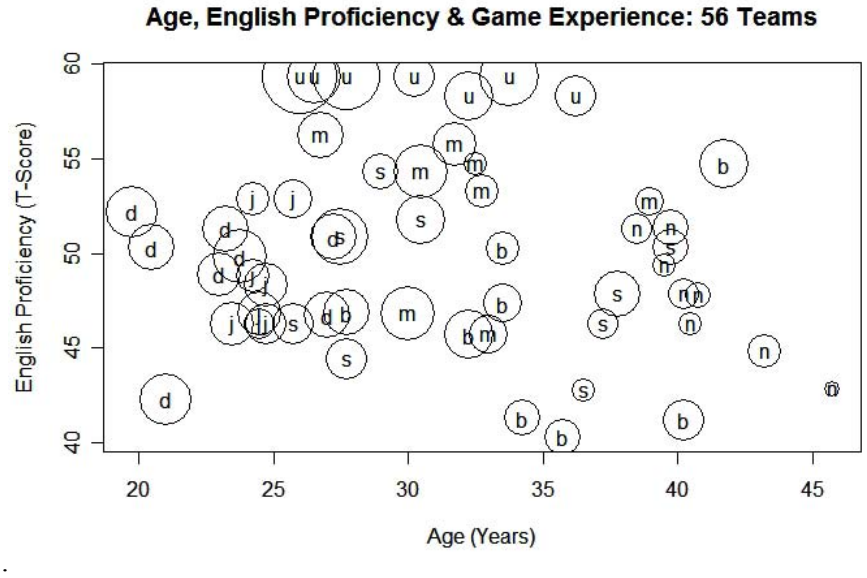

Figure 1: Demographic profiles of the 56 teams. Game experience is proportional to bubble size. Letters indicate national composition of the teams: Bulgaria (b), The Netherlands (d), Norway-senior age (n), Norway-junior age (j), Sweden (s), United States (u), mixed culture (m).

1.3 Procedure & Metrics

Each team member was seated at a computer terminal. Same-nation players were visually and auditorially shielded from the others at a site in their home nation, mixed-nation players played in their own nation over the Internet. After two to three hours of training and a break, players were briefed on their mission and engaged in a planning session before actual game-play proper. The main team task was to amass as many 'goodwill' points as possible. Points were primarily earned by finding weapons caches and lost by angering the local populace or by opening empty crates.

Since game-play and communication was by keyboard, every keystroke was available for analysis. During the game, there were probes from a "superior officer" to determine situation awareness at three different times. The primary dependent variable was the amount of goodwill points earned by the team. (Since team members could specialize such that a communications officer would not find any caches and a sensor operator might find several, individual scores are meaningless.) Other performance measures include the amount of communications and the degree of situation awareness.

1.4 Selected Results

Figure 2 shows the value of each team on the main performance metric (T-score of goodwill points) grouped by team national composition. It is clear that the mixed-nation teams are mostly in the upper-half of the performance distribution contrary to the hypothesis.

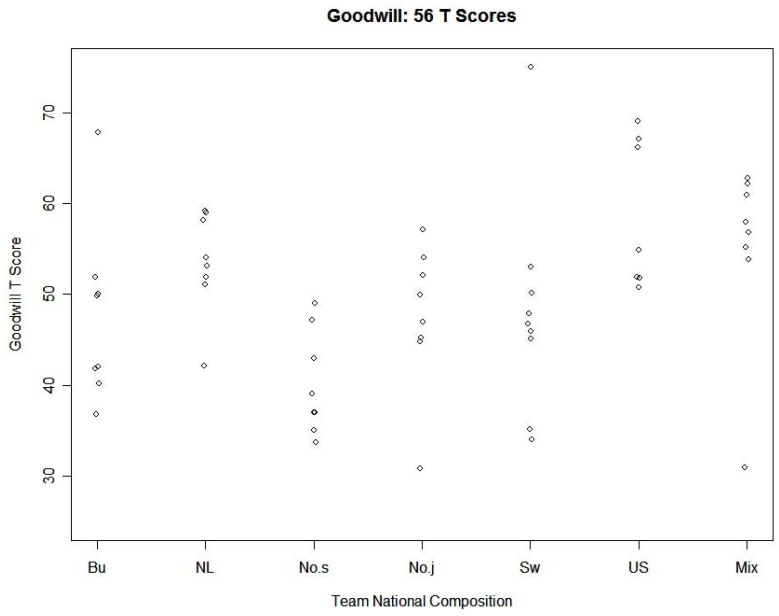

Figure 2: Team "goodwill" performance T-scores (Mean = 50; SD = 10) for each of the 56 teams grouped by national composition. Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age (No.s), Norway-junior age (NO.j), Sweden (Sw), United States (US), Mixed-nation (Mix).

2.0 LESSONS LEARNED

As the previous section suggests, the NATO RTO HFM-138/RTG experiment is conceptually simple but very complex with respect to methodological aspects such as the role-play game itself, details of the scenario, and a team's task and options. The experiment was also complex logistically both within a session and throughout the entire experiment. In conducting the experiment, we learned numerous lessons within the broad categories of conception, the game itself, methodology, logistics & execution, and analysis.

2.1 Concepts, Hypotheses, & Theoretical Issues

This experiment used a complex computer game to study adaptability in multinational coalitions. This is appropriate due to the inherent and pronounced immersive quality of such games, but also due to the fact that tomorrow's military recruits are growing up playing more and more such games and developing computer and communications skills not typical of people from a generation ago. Questions about what make some teams more effective than others are difficult to answer in general, but differential computer experience adds a fresh and urgent dimension to these questions about team adaptability especially in multinational coalitions.

2.2 The Game & Its Characteristics

As stated in Section 1.1, the game we used is based on a complex, very absorbing, and immersive role-play game, *Neverwinter Nights*. Using this game, BBN Technologies developed a general-purpose research tool termed SABRE (Situation Authorable Research Environment) (Warren et al. 2004). At the request of NATO HFM-138, Leung, Diller, and Ferguson (2005) also (Warren et al. 2005) developed a modern search-for contraband scenario specifically tailored for this experiment. Both the general SABRE tool and the specific LTAMC scenario were extensively piloted and iteratively refined, and we learned numerous lessons in the development phase and the execution phase of the research.

- Features that permit creativity, variant behaviors: The game and task were chosen so as to permit a large degree of creativity and self-determination by the teams in how they would approach accomplishing their mission. But the more degrees of freedom given the teams and the more unstructured the task, the less control that the experimenters have and the harder it is to interpret the various results. It should be noted that the experimental scenario was relative "static" in that there were no surprises or major incidents occurring during the game-play. The use of non-briefed events could certainly be introduced into the game, but we chose not to do so to maintain a degree of comparability in the experiment.
- Main tasks versus side quests: Although the scenario was relatively "static" as just discussed, there were some opportunities for teams to engage in "side quests" (such as helping a non-player character computer-generated girl search for a lost pet) which could garner goodwill points but which would take time away from the main task. Such side-quests do add realism and permit opportunities for non-routine decision making.
- Experimenter's viewpoint and used and unused game features: From the experimenter's viewpoint, the game can be very rich in decision making opportunities. However, a particular team might decline various opportunities or not be very creative and thus, as the game unfolds, the game can evolve into something less rich because certain avenues are not explored.
- Player's viewpoint: By observing the players and from their comments after the experiment, it is clear that the game succeeded in being immersive and absorbing. Players did not report trying to figure-out what the experiment was about, but rather quickly became fully engaged in the task at hand.
- Experimenter interaction/intervention possibilities: The underlying game (*Neverwinter Nights*) has a "dungeon-master mode" feature in which a game-master (or an experimenter) can have an invisible "avatar" (i.e., personal representative character in the game) which can interact with the game environment and other characters. We only used this feature on the rare occasions when a human player's avatar got "stuck" in a wall (there are occasional glitches since the software is very complex) to free the avatar without the human's awareness. One lesson learned is to be prepared for such events and to know how to deal with them.

- A related lesson for future research is that the dungeon-master mode can be utilized to introduce some player-action-contingent events into the game-play. For example, a door could be closed (by the unseen dungeon master) thereby trapping the human player until they radio for rescue by another player. Such in-game or in-line modifications require active monitoring and in-game intervention by an experimenter, but the possibilities are intriguing.
- Underlying & unused game features: Since the underlying game permits many behaviors which are not needed or allowed in a particular scenario (i.e., casting spells), it is important to prevent their accidental use by, for example, disabling the right-mouse button.

2.3 Methodology

The game and LTAMC scenario we used is complex to learn and complex to play, but the permitted behaviors are manifold. This richness means that certain methodological aspects that are normally under an experimenter's complete control in a more traditional laboratory experiment are not-controlled or even un-controllable. Some methodological lessons we learned or special problems we encountered in conducting the study are:

2.3.1 Participants

- Incomparability of subject pools: When participants come from multiple countries, it is very difficult to be sure that the subject pools are comparable. For example, a junior officer in one country might be considered a student in a second country and hence not in the pool of the second country.
- Size of subject pool: In spite of the size of many militaries, the pool of available participants can be surprisingly small. Military officers, in particular, are busy people and often have critical jobs from which they cannot be spared for a block of 4 to 6 hours. When constraints are placed on the characteristics of an entire team, such as requiring a certain age range, the effective size of the pool can, and does, shrink drastically.
- Representativeness of participants to intended application: Military officers have specialized occupations and some of these are not interchangeable. A medical officer cannot be expected to perform the work of a pilot. When the pool of possible participants is small, allowance must be made to permit more people to qualify for the experiment. Unfortunately, this means that the relevance of the results to the target population could become compromised.
- Team formation: Within a country and within the same research site, some individuals might know each other and some might be strangers. But teams whose members have a common past history can be expected to function differently than teams whose members are strangers. A background question about prior knowledge of or experience with other team members should be included along with the demographic questions.
- Distributed "team" issues & considerations: When team members come from different geographic locations or even nations, there are special issues of team formation and identification with the team. This problem is compounded when the only interaction team members can have is via a keyboard. But, however cumbersome "introductions" and interactions might be among distributed teams, such teams are becoming more and more common.
- Non-player characters: The town populace was comprised of computer-generated "non-player characters" (NPCs). The avatars of the human players could interact with the NPCs via scripted question and answer sets. The NPCs were programmed to make a variety of responses such as

providing tips regarding the whereabouts of suspicious activity. But some NPCs could lie (i.e., they were programmed to provide false information). NPCs have great potential in general for research purposes. We see this area as needing more work, but one which can bring rich rewards especially as the NPCs take on theoretically-based or empirically-grounded personality and cultural characteristics. The number, content, and veracity of messages should be addressed by any researcher.

2.3.2 Experimental Design

- As discussed above, the pool of potential participants can be very small. Thus, it is imperative to use as efficient an experimental design as possible with respect to the number of necessary participants.
- The experiment must also be very efficient with respect to its time demands. Six hours makes it hard to get participants and also can be a strain on the participants. The total amount of time includes time for pre- and post-game questionnaires. These need to be kept to a minimum.
- Statistical Design, matched samples, controlled & uncontrolled variables: Another consequence of the limited subject pool is that there are few possibilities for matching subjects on extraneous variables or for assigning subjects to pre-specified levels in a factorial design on factors such as age. In a companion paper, Warren (2008) has argued that full experimenter control over all variables of interest in a complex experiment is not just difficult but actually impossible. However, this does not mean that the effects of the confounding variables such as computer-game experience, English proficiency, or other covariates cannot be assessed. Using analysis of covariance (ANCOVA) and other regression-based techniques, these effects can be measured and then be statistically partialled-out.

2.3.3 Procedures

- The use of a computer game does not obviate the use of more traditional 5- or 7-point rating scales. We used both pre- and post-game questionnaires for obtaining such information as demographic data and personality and cultural profiles. But another feature that recommends use of the game is the occurrence of in-game probes. As mentioned earlier, on three occasions, a "superior officer" (wholly within the game), probed the participants with questions relating to their situation awareness. The use of in-game probes can be a powerful tool and is a supplement to the out-of-game questionnaires and the in-game situations (which are themselves tests).
- Training: different learning curves and times: There were two training phases, one in which individuals learned basic one-person actions such as moving forward, picking up objects, using a map, using one's journal, etc., and a second phase in which an individual learned to communicate with others. People were permitted to complete basic individual-action training at their own pace. But this meant that people finished basic training at different times. Fast learners often had to wait a long while at an in-game waiting area while slower learners were still mastering basic skills. The in-game waiting area had amusing activities to keep people busy, but it could be a long time, and the amusement nature of the filler activities could contribute to a sense that the overall game was not a serious exercise.
- Training: proficiency criteria and removal concerns: Related to the problem of different people taking time to reach a sufficient level of proficiency is the question at what level to set the proficiency criteria. Although this never occurred in the main experiment, we did have a case during piloting with non-military participants when one individual simply could not achieve sufficient skill to enable that experimental run to continue. Since this occurred during piloting, no time limit had been set, and this

led to a boredom problem with the other three players. Of course, not only do such aborted sessions waste peoples' time, it can be costly in terms of money since (non-military) participants still have to be paid.

- Local testing issues: breaks etc. When testing was at one site, the procedure was to conduct pre-questionnaire completion, individual, and team testing phases before lunch. The planning and search phases were after lunch, but this raises the chances that some forgetting might take place. We now recommend that a short "refresher" training session occur after lunch.
- Distributed testing: time zones consideration: The mixed-nation testing was done over the Internet. But since the experiment spanned 6 time zones and could take 6 hours, the experiment began relatively early in the morning for the Americans and ended relatively late at night for the Europeans. The previous point's reference to "lunch" has to be modified, but the issue of the timing of breaks becomes even more important. Anything that lengthens the experiment, such as the above recommendation for a "refresher" training phase, must be carefully weighed against the effect of a long day on some people's performance.

2.4 Administration & Logistics

- Subject scheduling issues: As discussed above, the size of the pool of possible participants was severely limited. One administrative difficulty that resulted from this was that of being able to schedule at least four people for a test day. It often took considerable effort on the part of the research team to locate and enlist the minimum of four people needed for a team.
- The difficulties were great enough that there were times when a session had to be canceled in advance due to either the inability to locate four participants or due to the advance cancellation by one of the volunteers. This again put a burden on the research team to contact the remaining volunteers.
- Even on days when four people had been scheduled, there was the all too common and exacerbating problem of a scheduled volunteer not appearing and thus forcing the cancellation of the session and the attendant loss of time of those who did appear for the experiment.
    - One technique for dealing with the problem of "no-shows" is to schedule more people than required. Due to our limited participant pool, this option was difficult to exercise.
    - Even if we had a large pool and could routinely "overbook" participants, overbooking does not guarantee that the required number of participants will show up. The reality of research on teams is that no-shows are all too common: If 6 people are scheduled for a four-person session, only three might show up.
    - But "overbooking" has is own problems. One problem is that if all show up for the experiment, some method has to be used to determine whom to dismiss and in such a way that the excused person is treated with respect and made to feel that their effort is still appreciated and not wasted.
    - In research without the need for a participants with highly specialized characteristics, one way to not waste any "unusable" participants who report for an experiment (either too few or too many) is to have alternate lower-priority experiments ready which can use whatever number of participants are available after due consideration for the needs of the highest priority experiment. However, this was not an option for us due to the small size of the pool of participants. Any potential participants who could not be run even after they reported for the experiment needed to be asked to reschedule if at all possible.

- Scheduling a long experiment over an ocean: The mixed team portion of the experiment often required having an American and Bulgarian on the same team. Arranging for a short meeting across 5 or more time zones is hard, but arranging a an experimental session that will take six or more hours means that the Europeans will be finishing quite late in their day and that the Americans will be starting quite early in their day. The definition of "lunch" break is thus relative and has to be taken into account when the potential participants are given details about what is being asked of them when they are solicited.
- Computer operators and local administrators: The above remarks about long experiments across an ocean also apply to the local computer operators and local experiment administrators who, by the nature of their responsibilities, must be present both before and after the participant session.
- Internet operators: The mixed team portion of the experiment also required the use of a knowledgeable team of SABRE experts and an internet operations center to "host" and coordinate the multi-site internet portion of the experiment. In order to ensure smooth operations and prevent loss of precious data, the internet operations had to be flawless. This required much advance preparation and testing of communication links and procedures. Although given scant mention in the experimental write-up and methods sections of the reports, this aspect of the experiment is crucial and required considerable effort.

2.5 Data Collection, Processing & Analysis

The SABRE testbed features automatic data collection of both pre-game questionnaires and within-game activity and communications. SABRE also collates the data from the various individual team sessions and collates the data into large spreadsheet files for post-processing by various statistical packages.

- Although SABRE does provide some basic statistics, it was felt best to leave the main analyses to the various members of the experimental teams and the statistical packages they prefer. One reason for this is the large and diverse nature of the data recorded and the subsequent opportunities for post-experiment data mining. We believe that the datasets resulting from this experiment will yield rich treasures as we continue to mine them.
- With a data set resulting from the game-play and questionnaires of 224 participants, it is invariable that there will be some missing data. Since different analysts have different preferences for dealing with missing data, it is imperative that there be tight configuration management of the raw and early-processed data sets that are distributed to the various analysts. In turn, it is also important that the various analysts maintain their own processed-data file configuration management with full description of the decisions they made and the procedures they followed.

2.6 Drawing Conclusions & Making Recommendation

In spite of running 224 participants, the resulting number of four-person teams was 56.

- Since our analyses are all team-centric, the conclusions are based on the relatively small number of 56 teams. As such, statistical power is weak and the conclusions must be taken with caution.
- Also, as discussed by Warren (2008), there are several confounds that also serve to temper our conclusions and recommendation such participant differences in age, computer-game experience, and English proficiency.
- However, the confounds are, to a large degree, unavoidable due to the complex nature of the participant populations. They are not deficiencies in the experimental design. Fortunately, there are

statistical techniques such as ANCOVA and linear regression which can "partial out" the effects of the confounds and enable the drawing of confound-free conclusions.

3.0 FINAL COMMENTS & GENERAL QUESTIONS

We have partial answers to what makes some teams perform well and others not so well. But, in general, much of what makes a team adaptable in a multinational coalition is still not fully understood. However, we believe we have demonstrated the value of using an immersive computer game to provide rich data sets to help provide such answers. As tomorrow's military recruits become more and more experienced with complex immersive computer games than the recruits of yesterday, it becomes imperative that we study the possible impact of such experience on selection and training for tomorrow's more computer-reliant military.

4.0 REFERENCES

[1] Leung, A., Diller, D., & Ferguson, W. (2005). SABRE: A game-based testbed for studying team behavior. *Proceedings of the Fall Simulation Interoperability Workshop (SISO*). Orlando, FL, September 18-23, 2005.

[2] NATO RTO HFM-138/RTG (2008). Report of the NATO Research and Technology Organization (RTO) Human Factors and Medicine Panel Research Task Group 138 on "Adaptability in Multinational Coalitions." Brussels: NATO.

[3] Stevens, J.P. (2002). *Applied multivariate statistics for the social sciences*, (4th Ed.). Mahwah, NJ: Erlbaum.

[4] Surowiecki, J. (2004/2005). *The wisdom of crowds.* New York: Anchor Books.

[5] Venables, W.N., & Ripley, B.D. (2002). *Modern applied statistics with S*, (4th Ed.). New York: Springer.

[6] Warren, R., Sutton, J., Diller, D., Ferguson, W., & Leung, A. (2004). A game-based testbed for culture & personality research. *Proceedings of the NATO Modeling and Simulation Group - 037 Workshop: Exploiting Commercial Games for Military Use*. The Hague, The Netherlands, 20-21 Oct 2004.

[7] Warren, R., Diller, D.E., Leung, A., Ferguson, W., & Sutton, J.L. (2005). Simulating scenarios for research on culture & cognition using a commercial role-play game. In M.E. Kohl, N.M Steiger, F.B. Armstrong, and J.A. Jones, (Eds.), *Proceedings of the 2005 Winter Simulation Conference*. Orlando, FL.